Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 11:30:19
REFERENCE NO: 232

Author Names & Affiliations

- Juan Zhao - Department of Electrical and Computer Engineering, Tennessee State University
- Suping Zhou - Department of Agricultural and Environmental Sciences, Tennessee State University

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Dr. Juan Zhao focus on machine learning, data mining and cloud computing. Dr. Suping Zhou focuses on Genetics, Physiology and Molecular Biology of Abiotic Stress in Plants.

Title of Submission

Towards Real Time Streaming Data Analytics

Abstract (maximum ~200 words).

The increased interconnectivity between the physical and cyber systems has resulted in torrents of streaming data. Analysis of the streaming data will provide key insights required for decision making and prediction. However, the dynamic changes in the physical and cyber system necessitate analytics to provide real-time insights. Real-time data analytics presents incredible opportunities in areas such as terrorism alerts, traffic navigation, security incident response, but also poses great challenges for the current technologies and techniques. We provide many challenges for real-time analytics that we are facing in conducting research in the smart- agriculture. The challenges for real-time analytics mainly lies in models generation/update/prediction and ability to handle massive data with unprecedented speed in real time. Thus, there is an urgent need to build a more effective Cyber-infrastructure to handle and facility the process of real-time data, so that it can contribute to the science research.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Currently, with the rapid development of Internet-of-Things (IoT), data can arrive in multiple, continuous, rapid and time-varying data streams. Analytics of the real-time data are becoming crucial in the study of problems such as terrorism alerts, self-driving, disease prediction, weather prediction and smart agriculture. However, handling and processing these real-time data pose great challenges for conduct research effectively, which mainly lies in three aspects: Challenges mainly lie in three aspects:
1) Data aggregation from geographically disparate real-time sources. Take our recent research activity on agriculture as an example. When we study environmental parameters and the microbial contamination to plants, we need to monitor and predict the plant growth by

collecting and analyzing real-time data from multi-sources, such as temperature and humidity of the soil sensors, and pictures taken from the plants and fields by cameras and drones. Data may arrive at a different time or contain missing values. Aggregating those data into each entity at real-time is quite challenged, and 2) High volume data need a scalable data mining and machine learn algorithms to reduce the computing complexity, and 3) In real time scenarios, the granularity of concept drift varies due to unpredictable changes in system dynamics, which poses great challenges for current learning algorithms to inclemently update models to handle the concept drift.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

Modern cyberinfrastructure, which is enabled by distributed computing and cloud techniques, has facilitated big data storage and processing. However, there is still strong requirements for current cyberinfrastructure to facility the real-time data process.
• To address data aggregation challenge, there is a need for developing real-time cyberinfrastructure, not only facilitated by High-speed wireless sensor network techniques to make sure high-quality data transmitted in real-time, but also sensor data processing software to assure ordered and timely delivery of messages to ensure models reflect the timely system dynamics.
• To address huge data processing and mining challenges, there is a need for fast, scalable and fault tolerant infrastructure to enable computation of model generation and prediction on the massive and high speed streaming data stream with zero latency.
• To address the concept drift challenges, there is a need for developing low false alarmed Event Detection Modules that can handle the concept drift. Also, need to build a bundle of enhanced machine learning Libs which can incrementally update the models.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

To build an effective Cyberinfrastructure to facility cross- disciplinary scientist to conduct research, we also need to consider the following issues:
• There is still a big gap between domain scientist and data scientists to produce science together. Thus it highly needs to develop a new generation of scientific data experts and scientific software engineers who can interact with science domain experts.
• Provide significant compute infrastructure for managing, analyzing and simulating the data generated by the facilities and for designing next generation Big-Science experiments.
• Increases capacity in advanced software development for data analysis and interpretation.

Consent Statement